EFFECT OF NON-NORMALITY ON THE ESTIMATION OF FUNCTINNS OF VARIANCE COMPOOENTS

By

R.A. Singhal

Indian Veterinary Research Institute, Izatnagar-243 122 (U.P.)

(Received : May, 1981)

Summary

Approximate expressions for the mean value and variance of estimators of functions of variance components in balanced one-way random model (Model II) in samples from non-normal populations are derived and found to contain corrective-terms due to finite cumulants in addition to normal theory terms. The mean values are fairly insensitive but variances highly sensitive to the non-normality in random effects.

I. Introduction

Problems relating to the estimation of variance components and their functions, under normal theory assumption, have been discussed, among others, by Daniels [4], Kempthorne [8], Graybill [6] and Searle [13]. Considering samples from non-normal populations characterized by cumulants Hemmersley [7] and Tukey [16] derived expressions for the variances of variance components estimators in different designs and found them to contain corrective terms in population kurtosis of random effects in addition to normal theory terms.

Atiqullah [1], asymptotically, derived confidence limits for a ratio of variance components and found them insensitive to the kurtosis of random effects. Scheffee [12] and Bay [2] investigated effects of parental non-normality on the sampling distributions of functions of variance components and inferred that the distribution would be robust against the violation of normality assumption in error effects but sensitive to the non-normality in group effects.

Here approximate expressions for the first two moments of the estimators of different functions of variance components are obtained in order to investigate effects of departure from normality in random effects in a balanced one-way random model. An illustrative numerical example in section 4 characterizes the effects of non-normality.

2. MODEL AND VARIANCES AND COVARIANCES OF VARIANCE COMPONENTS ESTIMATORS

Suppose an observation y_{ij}, the jth ($j=1, 2, \ldots, r$) replicate in the ith ($i=1, 2, \ldots, g$) group, is represented as

$$y_{ij}=\mu+a_i+e_{ij}. \qquad \ldots(2.1)$$

Here μ is the overall mean, a_i is the ith group (random) effect with mean zero and variance σ_a^2 and e_{ij} is the residual effect with mean zero and variance σ_e^2. The σ_a^2 and σ_e^2 are known as variance components associated with group and error effects in the balanced one-way random effects model (2.1). We further assume that a_i and e_{ij} are identically and independently distributed and are random samples from infinite populations represented by the first four terms of Edgeworth series with third and fourth cumulants $\lambda_{3_a}(=\sqrt{\beta_{1_a}})$ and $\lambda_{4_a}(=\beta_{2_a}-3)$, and λ_{3_e} and λ_{4_e}, respectively.

The 'between-groups' and 'within-groups' mean squares are

$$s_g^2=\sum_{i=1}^{g}(\bar{y}_{i.}-\bar{y}_{..})^2/p$$

and

$$s_e^2=\sum_{i=1}^{g}\sum_{j=1}^{r}(\bar{y}_{ij}-\bar{y}_{i.})^2/q$$

with $p=(g-1)$ and $q=g(r-1)$ degrees of freedom, respectiveiy. The estimators of σ_a^2 and σ_e^2, in analysis of variance, are

$$\hat{\sigma}_a^2=(s_g^2-s_e^2)/r \qquad \ldots(2.2)$$

and

$$\hat{\sigma}_e^2=s_e^2, \qquad \ldots(2.3)$$

respectively.

Deriving joint sampling distribution of s_g^2 and s_e^2 in samples from non-normal populations referred to above and whence the exact sampling distributions of $\hat{\sigma}_a^2$ and $\hat{\sigma}_e^2$ Singhal [15] obtained by direct evaluation that

$$E(\hat{\sigma}_a^2)=\sigma_a^2,\ E(\hat{\sigma}_e^2)=\sigma_e^2, \quad ...(2.4)$$

$$V(\hat{\sigma}_a^2)=\frac{2}{r^2}\left(\frac{(r\,\sigma_a^2+\sigma_e^2)^2}{p}+\frac{\sigma_e^4}{q}\right)+\frac{\sigma_a^4\lambda_{4a}}{g}, \quad ...(2.5)$$

$$V(\hat{\sigma}_e^2)=\frac{2\,\sigma_e^4}{q}+\frac{\sigma_e^4\,\lambda_{4e}}{N}, \quad ...(2.6)$$

and

$$Cov\,(\hat{\sigma}_a^2,\hat{\sigma}_e^2)=-\frac{2\,\sigma_e^4}{N}. \quad ...(2.7)$$

Where $N=rg$ and E (.), V (.) and Cov (.,.) are mean value, variance and covariance, respectively.

Surprisingly these expressions though obtained for samples from Edgeworth population which assumes a positive definite and unimodel density function within Barton & Dennis [3] and Singh [14] limits $(-1.0\leqslant\lambda_4\leqslant 2.4,\ 0\leqslant\lambda_3^2\leqslant 0.5)$ but are found in exact agreement with Hemmersley [7], Tukey [16] and Atiqullah [1] which they achieved differently for samples from populations specified by cumulants of random effects.

Further, it is easy to show that

$$Cov\,(\hat{\sigma}_a^2+\hat{\sigma}_e^2,\ \hat{\sigma}_a^2)=\frac{2\,\sigma_a^4}{p}+\frac{4\,\sigma_a^2\,\sigma_e^2}{rp}+\frac{2\,\sigma_e^4}{Nrp}+\frac{\sigma_a^4\,\lambda_{4a}}{g} \quad ...(2.8)$$

and

$$Cov\,(\hat{\sigma}_a^2+\hat{\sigma}_e^2,\ \hat{\sigma}_e^2)=\frac{2\,\sigma_e^4}{N}+\frac{\sigma_e^4\,\lambda_{4e}}{N}. \quad ...(2.9)$$

It is seen that the derived expressions are quadratic functions of the variance components being estimated and contain corrective terms due to finite cumulants in addition to normal theory terms.

3. Estimators of Functions and their Variances

In applied work such as quantitative genetics, breeding, industrial statistics, psychology and sample survey we come across with the problem of estimation of parameters which are functions of variance components σ_a^2 and σ_e^2. Some functions of interest are:

The variance of tte observation y_{ij}, *i.e.*

$$V_p=\sigma_a^2+\sigma_e^2, \quad ...(3.1)$$

is called the total variance or the phenotypic variance in genetics and breeding. The correlation among the observations of a group, *i.e.*

$$t=\sigma_a^2/(\sigma_a^2+\sigma_e^2), \quad ..(3.2)$$

is called the intraclass correlation (Fisher, [5] and is commonly used by geneticists and animal breeders in studies relating to heritability, *i.e.*

$$h^2=4\sigma_a^2/(\sigma_a^2+\sigma_e^2). \quad ...(3.3)$$

The range of variation for t is from $-(r-1)^{-1}$ to 1 and for h^2 from 0 to 1 instead of $-4/(r-1)$ to 4, which we encounter in practice. The other functions of interest are

$$R=\sigma_e^2/(\sigma_a^2+\sigma_e^2),\ 0\leqslant R\leqslant\frac{r}{(r-1)} \quad ...(3.4)$$

and

$$W=\sigma_a^2/(\sigma_e^2,\ 0\leqslant W\leqslant\infty. \quad ...(3.5)$$

The estimators of these parameters are

$$\hat{V}_p=\hat{\sigma}_a^2/+\hat{\sigma}_e^2, \quad ...(3.6)$$

$$\hat{t}=\hat{\sigma}_a^2/(\hat{\sigma}_a^2+\hat{\sigma}_e^2), \quad ...(3.7)$$

$$\hat{h}^2=4\hat{\sigma}_a^2/(\hat{\sigma}_a^2+\hat{\sigma}_e^2), \quad ...(3.8)$$

$$\hat{R}=\hat{\sigma}_e^2/(\hat{\sigma}_a^2+\hat{\sigma}_e^2), \quad ...(3.9)$$

and

$$\hat{W}=\hat{\sigma}_a^2/\hat{\sigma}_c^2, \quad ...(3.10)$$

respectively.

Since these estimators are functions of the unbiased estimators $\hat{\sigma}_a^2$ and $\hat{\sigma}_e^2$ we investigate, in what follows, the unbiasedness of these estimators.

3.1 Unbiasedness of Estimators

The unbiasedness property of the estimators (3.6) to (3.10) is studied using the expressions (2.4) to (2.9) into an approximate formula of Rao ([10], p 154) for the ratio of two unbiased estimators. After algebraic simplifications, we get

$$E(\hat{V}_p) = \sigma_a^2 + \sigma_e^2,$$

$$E(\hat{t}) = t + b(t),$$

$$E(\hat{h}^2) = h^2 + b(h^2) = h^2 + 4b(t),$$

$$E(\hat{R}) = R + b(R) = R - b(t),$$

Where

$$b(t) = -\frac{2t^2(1-t)}{p} - \frac{2(1-t)^2((N-r+1)t+1)}{Nrp}$$
$$-t(1-t)\left(\frac{t\lambda_{4a}}{g} - \frac{(1-t)\lambda_{4e}}{N}\right)$$

and for sufficiently large r

$$b(t) \doteq -\frac{2t^2(1-t)}{p} - t(1-t)\left(\frac{t\lambda_{4a}}{g} - \frac{(1-t)\lambda_{4e}}{N}\right). \quad \ldots(3.11)$$

The symbol '$\doteq$' denotes approximately equal. Apparently $b(t)$, the bias, is a function of t and is always in the negative direction. It tends to zero as t approaches the extreme ends. However, for moderate t, g and r bias is likely to contribute appreciably in the estimation of the parameters. Considering samples from normal population Fisher [5] used $z = \frac{1}{2}\log_e\,(1+(r-1)\hat{t})/(1-\hat{t})$ transformation which tends to normality for increasing g and suggested to add the correlation factor $+\frac{1}{2}\log_e\,(g/(g-1)$ or approximately

$$+(2g-1)^{-1}.$$

Lastly

$$E(\hat{W}) = w + b(w),$$

where

$$b(w) = \frac{2}{rq} + \frac{2w}{q}\left(1 + \frac{q\lambda_{4e}}{N}\right). \quad (3.12)$$

Graybill [6] approached the unbiased estimator of w from variance ratio under normal theory which deviates from (3.12), if $\lambda_{4e} = 0$, on account of approximation but the difference would be negligible.

It is seen that the estimator (3.6) is unbiased whereas the estimators (3.7) to (3.10) are biased and the bias can be removed by using correction term b (.). The estimates of bias can be computed and will provide rough, but useful, check on the size of bias in a specific sample.

3.2 Sampling Variances of Estimators

Sampling variances of the estimators (3.6) to (3.10) are obtained utilizing (2.4) to (2.9) into the approximate formula (10.17) of Kendall & Stuart [9]. After heavy algebraical simplification we obtain

$$V(\hat{V}_p)=\frac{2\sigma_a^4}{p}+\frac{4\sigma_e^2\,\sigma_a^2}{r_p}+\frac{2((N-r+1)r-1)\,\sigma_e^4}{Npr^2}$$

$$+\frac{\sigma_a^4\,\lambda_{4a}}{g}+\frac{\sigma_e^4\,\lambda_{4e}}{N} \quad \ldots(3.13)$$

$$\doteq\frac{2\sigma_a^4}{p}+\frac{4\sigma_e^2\,\sigma_a^2}{r_p}+\frac{2\sigma_e^4}{N}+\frac{\sigma_a^4\,\lambda_{4a}}{g}+\frac{\sigma_e^4\,\lambda_{4e}}{N}, \quad \ldots(3.14)$$

$$V(\hat{t})=\frac{2(1-t)^2\,(1+(r-1)\,t)^2\,(N-1)}{pqr^2}$$

$$+t^2(1-t)^2/\left(\frac{\lambda_{4a}}{g}+\frac{\lambda_{4e}}{N}\right) \quad \ldots(3.15)$$

$$\doteq\frac{2(1-t)^2\,(1+(r-1)\,t)^2}{r(r-1)g}+t^2(1-t)^2\left(\frac{\lambda_{4a}}{g}+\frac{\lambda_{4e}}{N}\right), \quad \ldots(3.16)$$

$$V(\hat{h}^2)=16V(\hat{t}) \quad \ldots(3.17)$$

$$V(\hat{R})=V(\hat{t}) \quad \ldots(3.18)$$

and

$$V(\hat{w})=2\left(\frac{1}{p}+\frac{1}{q}\right)\left(w+\frac{1}{r}\right)^2+w^2\left(\frac{\lambda_{4a}}{g}+\frac{\lambda_{4e}}{N}\right) \quad \ldots(3.19)$$

As a check, it is found that expressions (3.16) and (3.19) under normal theory, exactly agree with Fisher [5] and Scheffee [12]. Fisher while discussing the limitations of $V(\hat{t})$ has suggested that it should not be used for testing the significance.

Above results suggest that the sampling distributions of the estimators would be little affected by the violation of normality in 'error effects' but would be appreciably affected by the presence of kurtosis in 'group effects' on account of the multiplying factors $1/g$ and $1/N$ and it is also expected in practice that the 'error effects' are near normal than 'group effects'.

For immediate practical application we compute

$$qs_e^4 \,/\, (q+2) \qquad \ldots(3.20)$$

and

$$p'(s_g^2 - s_e^2)^2/(p'+2)r^2, \qquad \ldots(3.21)$$

where $p'=(s_g^2 - s_e^2)^2/\,(s_g^4/p + s_e^4\,/q)$ is approximate degrees of freedom for σ_a^2 (Satterthwaite, [11]), as estimators of σ_e^4 and σ_a^4, respectively. The estimators of λ_{4a} and λ_{4e} are k_{4a}/k_{2a}^2 and k_{4a}/k_{2e}^2, respectively, where k_{4a} and k_{2a} and k_{4e} and k_{2e} are Fisher's k-statistics computed from the group means and from the pooled within group variation.

Thus thc estimators of $V(\hat{\sigma}_e^2)$ and $V(\hat{\sigma}_a^2)$ are

$$\hat{V}(\sigma_e^2) = \frac{2s_e^4}{(q+2)}\left(1 + \frac{q\hat{\lambda}_{4e}}{2N}\right) \qquad \ldots(3.22)$$

and

$$\hat{V}(\hat{\sigma}_a^2) = \frac{2}{r^2}\left(\frac{s_g^4}{(p+2)} = \frac{s_e^4}{(q+2)}\right) + \frac{\hat{\sigma}_a^4\,\hat{\lambda}_{4a}}{2g}\,, \qquad \ldots(3.23)$$

since it is well known that $E\,(s_g^2) = r\ \sigma_a^2 + \sigma_e^2$ and $E\,(s_e^2) = \sigma_e^2$.

The estimates of $b(.)$ and $V(.)$ can be obtained using estimates of parameters for specific sample studies.

4. A Numerical Example

The birth weights (kg) of $N=75$ progenies of $g=15$ Hoiltein Friesian sires (male parents) each mated over a period of years to $r=5$ randomly selected Hariana females under a cross-breeding programme at Indian Veterinary Research Institute, Izatnagar were considered. The analysis of this set of data gives :

$$s_g^2 = 4299.40,\ s_e^2 = 2333.90 = \hat{\sigma}_e^2,\ \hat{\sigma}_a^2 = 393.10,$$

$$\hat{\sigma}_e^4 = 5271376.65,\ \hat{\sigma}_a^4 = 89293.06,\ \hat{\lambda}_{4a} = -1.1394,$$

$$\hat{\lambda}_{4e} = -0.3399,\ p=14,\ q=60 \text{ and } p'=2.74$$

TABLE 1

Comparative values for the estimates of parameters and their variances

Estimator	*Conveutional estimates*	*Correction*		*Unbiased estimate*		*Variances*		
		Normal	*Non-normal*	*Normal*	*Non-normal*	*Correction*	*Normal*	*Non-normal*
$\hat{\sigma}_e^2$	2333.9009	—	—	2333.9000	2333.9000	—23889.8789	175712.5551 (419.1808)	151822.6762 (389.6442)
$\hat{\sigma}_a^2$	393.1000	—	—	393.1000	393.1000	—6782.7005	99452.7040 (315.3612)	92670.0935 (304.4174)
$\hat{V}_p$	2727.0000	—	—	2727.0000	2727.0000	—30672.5794	205752,2572 (453.5992)	175079.6778 (418.4252)
$\hat{t}$	0.1442	—0.0025	0.0009	0.1459	0.1459	—0.0012	0.0130 (0.1140)	0.0118 (0.1085)
$\hat{h}^2$	0.5756	—0.0102	0.0035	0.5868	0.5833	—0.0196	0.2081 (0.4562)	0.1885 (0.4341)
$\hat{R}$	0.8558	0.0025	—0.0009	0.8533	0.8541	—0.0012	0.0130 (0.1140)	0.0118 (0.1085)
$\hat{\omega}$	0.1684	0.0123	—0.0008	0.1561	0.1569	—0.0023	0.0239 (0.1546)	0.0216 (0.1471)

Figures (.) denote standard error.

The conventional estimates and variances of estimates along with corrections are tabulated in the Table. The corrections for bias computed using approximate expression for $b(.)$ suggest that the magnitude of correction is more for normality than for non-normality. Thus the estimates of the ratios of variance components, in this example, are not seriously affected by the non-normality in random effects. From the results it is found that the variances of the functions are very sensitive to changes in the population form. In all the cases the actual variance (standard error) is found to be widely divergent from its normal theory value which would effectively alter inferences about functions of variance components.

ACKNOWLEDGEMENTS

Author is grateful to Dr. C. Singh for constant advice in this work and to referee for helpful comments.

REFERENCES

[1] Atiqullah, M. (1962) : On the effect of non-normality on the estimation of components of variance. *J. Roy. Statist. Soc.* **B 24**, 140-47.

[2] Bay, K.S. (1973) : The effect of non-normality on the sampling distribution and standard error of reliability coefficient estimates under an analysis of variance model, *Brit. J. Math. Statist. Psychol.* **26**, 45-57.

[3] Barton, D.E. and Dennis, K.E. (1952) : The conditions under which Gram-Charlier and Edgeworth curves are positive definite and unimodal. *Biometrika*, **39**, 425-7.

[4] Daniels, H.E. (1939) : The estimation of components of variance. *J. Roy. Statist Soc. Suppl.* **6**, 186-97.

[5] Fisher, R.A. (1925) : *Statistical Methods for Research Workers.* Oliver & Boyd, London.

[6] Graybill, F.A. (1961) : *An Introduction to Linear Statistical Models.* Vol. I, McGraw Hill, pp. 337-82.

[7] Hemmersley, I.M. (1949) : The unbiased estimate and standard error of the intraclass variance, *Metron*, **15**, 189-205.

[8] Kempthorne, O. (1957): *An Introduction of Genetic Statistics*, pp. 129-85, Wiley, New York.

[9] Kendall, M.G. and Stuart, A. (1963) : *The Advanced Theory of Statistics.* Vol. I, 2nd edn, Charles Griffin & Co., London.

[10] Rao, C.R. (1957) : *Advanced Statistical Methods in Biometric Research*, Wiley, New York.

[11] Satterthwaite, F.F. (1946) : An approximate distribution of estimates of variance components. *Biometrics Bull.* 2, 110-14.

[12] Scheffee, H. (1959) : *The Analysis of Variance*. Wiley, New York.

[13] Searle, S.R. (1971) : Topics in variance components estimation. *Biometrics*, 27, 1-76.

[14] Singh, C. (1970) : The Distribution of Range and Certain Related Statistics in Samples from Non-normal Populations. Ph. D. Thesis, University of Delhi.

[15] Singhal, R.A. (1983) : Effect of Non-normality on Some Tests in Analysis of Variance. Ph. D. Thesis, submitted to Rohilkhand University.

[16] Tukey, J.W. (1956) : Variances of variance components I. Balanced designs. *Ann. Math. Statist.*, 27, 722-36.